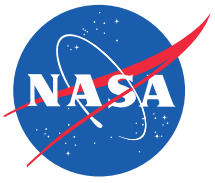
JPL Publication 13-12



Characterizing Interference in Radio Astronomy Observations through Active and Unsupervised Learning

G. Doran
Jet Propulsion Laboratory

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

October 2013

I would like to thank my mentor Kiri Wagstaff, co-mentor David Thompson, and the rest of the Machine Learning and Instrument Autonomy group for an excellent internship experience during Summer 2012. I would also like to thank the radio astronomers at JPL, Sarah Burke-Spolaor, Jake Hartman, Dayton Jones, Joseph Lazio, Walid Majid, and Robert Preston for interesting discussions, guidance, and feedback. Additionally, I would like to acknowledge collaborators at the NRAO, Paul Demorest, Mike McCarty, and Mark Whitehead for numerous conversations.

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Table of Contents

1 Introduction

In the process of collecting radio signals from astronomical sources, radio astronomers are faced with the challenge of mitigating the effects of radio frequency interference (RFI) resulting from man-made radio sources. Examples of such sources include cell phones, satellites, aircraft, and on-site observatory equipment [1]. As the volume of data collected and prevalence of RFI sources increases, automated mitigation of RFI becomes increasingly important.

Below, I describe machine learning approaches to automatic RFI detection and classification. As opposed to previous work, which tends to focus exclusively on RFI flagging or excision [1, 2, 3], this report describes techniques that aim to provide additional insight into various types and sources of RFI phenomena. The approaches investigated use unsupervised and active learning to minimize the need for human interaction and to allow large volumes of data to be processed efficiently.

Much of the observational data used for this project comes from the Parkes Multibeam Pulsar Survey. However, I discuss possibilities for extending the techniques explored to other instruments or observatories, such as the Green Bank Telescope.

2 Background

RFI mitigation involves the selection of a proper observatory location, enforcement of policies intended to reduce nearby use of radio devices, and shielding of on-site radio-emitting equipment [4]. However, RFI from sources such as satellites will always be present, requiring additional measures to filter interference. Therefore, RFI mitigation techniques are also applied at various points throughout the data collection process. In particular, *pre-correlation* filtering is performed at single receivers before the signals from multiple receivers have been correlated and integrated into a single signal. However, this requires an algorithm that can quickly process large amounts of data with high time resolution [3]. An alternative is to excise interference *post-correlation*, which allows for more sophisticated off-line data analysis [4].

Though post-correlation RFI excision is often done manually [4], numerous techniques exist to automatically flag or remove data thought to contain interference [1, 2, 3]. Many existing techniques use a threshold-based approach in the time-frequency representation of the data. Due to the various characteristics of RFI sources (e.g. short wide-band bursts and persistent narrow-band signals), different detection and mitigation strategies are proposed for each [2]. Accordingly, it is observed that there is no single, universal algorithmic solution to RFI excision [5].

My project helps to close the loop between automated RFI detection and excision, and mitigation in the field. By automatically classifying and characterizing detected RFI events, machine learning techniques can aid experts in identifying distinct RFI sources. Furthermore, knowing properties of events, such as the location of the source or likely times of occurrence, provides valuable information that can be used to formulate and prioritize mitigation strategies.

A system to detect and characterize RFI events must meet several real-world requirements. For example, techniques used must be scalable to large quantities of data. In particular, event detection procedures must be completed in near real-time to keep up with a steady stream of observational data. Furthermore, classification techniques must

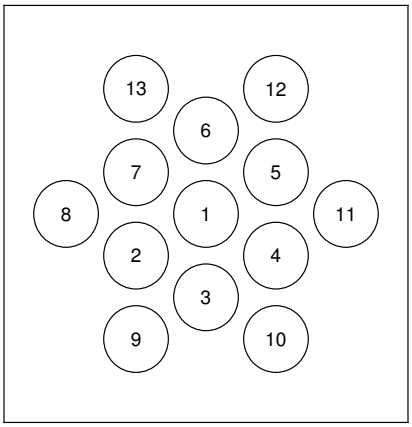


Figure 1: The configuration of the 13 beams (as projected onto the sky) of the Parkes Multibeam Receiver.

minimize the need for human interaction, since there are too many events to be manually labeled by an expert. Below, I describe an approach developed to achieve these goals.

3 Methods

3.1 Dataset

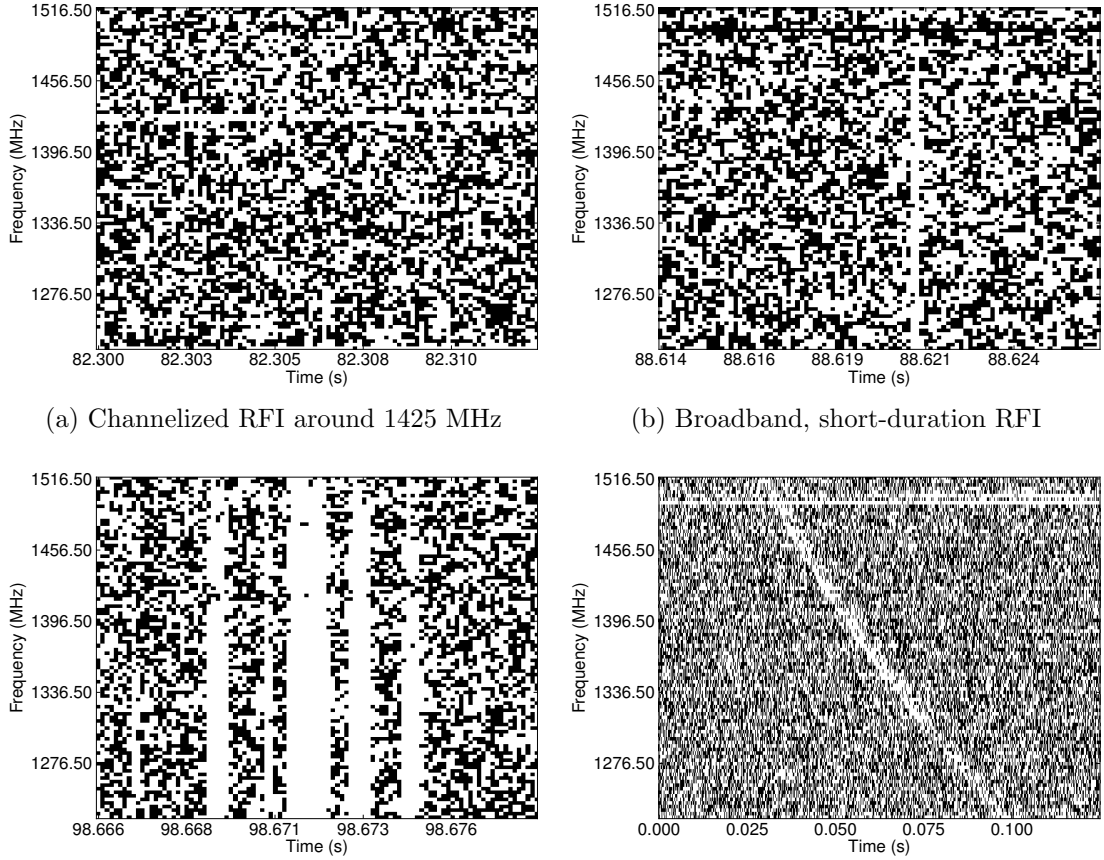
The data used for this project come from the Parkes Multibeam Pulsar Survey [6], carried out at the Parkes Observatory in Australia. The instrument used for observation receives 13 "beams" of signal, depicted in Figure 1, which correspond to distinct (though adjacent) portions of the sky [7]. Astronomical signals such as pulsars are usually only detected in between one to three adjacent beams. However, it is not uncommon for RFI to be present in all beams.

In the dataset, radio observations are represented as 1 bit intensity values in time-frequency space. Therefore, an observation can be thought of as a two-dimension image in which each pixel is either "on" or "off." There are approximately 1000 hours (3.5 TB) of observations, made over a period of 4.5 years from 1998 to 2002. The observations are split across about $12,000$ sets of 13 files for each beam.

The data from the Parkes Multibeam Pulsar Survey is represented as a *dynamic spectrum*, which consists of a spectrum of 96 frequency channels sampled every 125 μs. The frequency channels range from 1232.1 MHz to 1516.5 MHz spaced by 3 MHz. Some sample observations are shown in Figure 2.

RFI comes in two major varieties, channelized and broadband. Channelized RFI is typically present for long durations, but only in a few frequency channels. An example can be seen around 1425 MHz in Figure 2a. Broadband RFI corresponds to events that last for a short duration (on the order of milliseconds or less), but across many observed frequency channels (see Figure 2b). Occasionally, several broadband bursts will occur within a short time, such as in Figure 2c.

On the other hand, actual signals of interest, such as those from pulsars, tend to exhibit other characteristics due to *dispersion*. Dispersion occurs because lower-frequency signals travel more slowly through the interstellar medium than higher frequencies do [8], and leads to a signal that appears curved in the time-frequency plot by the time it reaches



(a) Channelized RFI around 1425 MHz

(b) Broadband, short-duration RFI

(c) An RFI event consisting of several broadband bursts

(d) A dispersed pulsar signal

Figure 2: Examples of RFI and astronomical signals within Parkes observations

the telescope (see Figure 2d).

3.2 RFI Detection

Because this project focuses on characterizing short, transient RFI events within Parkes observations, a simple thresholding method (such as described in prior work [2]) is used to identify samples in the time domain that are contaminated with RFI. First, the signal is averaged across all beams to highlight RFI that is observed in multiple beams. Then, for each time sample, the intensity values are averaged across all frequencies, and the sample is flagged as an event if the average exceeds some threshold (the threshold was set to 0.7 for this study). Flagged samples within 50 time steps (6.25 ms) of each other are temporally grouped together as a single "event." These parameters are arbitrary and can be adjusted as necessary. This simple detection procedure avoids detection of astronomical signals, which are dispersed by the interstellar medium. On the other hand, it efficiently captures many RFI events, which are not dispersed because their sources are located relatively close to the receiver.

3.3 Discriminating Features

There are several features that can be used to distinguish between different RFI phenomena. The first is the frequency structure of the event. Because there might be a complex relationship between the orientation of the telescope and how the RFI is observed in each beam, the intensities are averaged across beams to get a more stable view of which frequencies are most present in the RFI. For events that last more than a single time sample, the intensities can also be averaged across time to summarize the frequency structure of the event. As a result, there are 96 frequency features (one for each channel) with an average intensity value in the range $[0, 1]$.

Another feature that is useful for characterizing RFI sources is the time of day the event was observed. Since most RFI comes from human use of equipment, many RFI sources should be more active during the work day. Therefore, time of day was also used as a feature to discriminate between RFI events. To improve interpretability, all times are converted to local time.

Finally, another source of information about RFI events is the direction the telescope is pointing when they are observed. Due to shielding from the dish, the receiver is more likely to detect events in front of the telescope dish than behind it. The orientation of the telescope is described by azimuth angle and zenith angle. The azimuth angle is the compass angle clockwise from North where the telescope is pointing. The zenith angle is the angle the telescope makes with the line running through the center of the telescope perpendicular to the ground. The maximum zenith angle is 60°, since the large dish collides with the ground at higher zenith angles. One might also expect to detect more RFI near the horizon as the zenith angle increases.

Some care must be taken in representing the features for time and pointing direction. Because hours of the day are cyclical, 1:00 and 23:00 are closer, for example, than 1:00 and 6:00. However, simply using time of day as a feature means that points with times of 1:00 and 23:00 will have a larger Euclidean distance than those with times of 1:00 and 6:00, all else equal. Because algorithms like k-means (described below) use Euclidean distance to quantify the similarity of points, a flat feature representation for time is clearly undesirable. Therefore, times are represented as points on the unit circle in a two-dimensional subspace of the features space. More specifically, each time is represented using the sine and cosine of the angle that the hour hand of a 24-hour clock makes with 0:00. Similarly, the pointing angle is represented as a point on a three-dimensional hemisphere sitting atop the telescope. Combined with the frequency features, a total of 101 features are used to describe events.

3.4 Characterization

After RFI events have been described with appropriate features, there are several approaches that can be used to classify and categorize distinct RFI phenomena. Most standard machine learning algorithms work using *supervised learning*, in which labeled examples are used to train a classifier. Then, the trained classifier is used to label novel events that are observed. However, there is no known prior work on automatically categorizing transient RFI events, and therefore no source of labeled examples that can be used for training.

Alternatives to supervised learning include *unsupervised learning* and *active learning*. Unsupervised approaches attempt to classify events based on properties of the data distribution in the feature space. For example, one common assumption is that each data

point is drawn from one of k Gaussian distributions. The goal of algorithms such as k-means is to infer the most likely centers of the k distributions from which the data are drawn, given the observations [9].

To enable the scalability of unsupervised methods to handle millions of events, this project uses the *mini-batch* k-means algorithm [10]. Mini-batch k-means uses a stochastic gradient descent approach in which an approximate gradient computed from a small batch of examples is used to optimize the objective function for determining cluster centers.

Another approach used for this project is active learning, which seeks to learn an accurate classifier by querying a human expert for labels of the fewest number of examples possible [11]. At each step of an iterative process, an active learning algorithm selects an example about which it is most unsure and presents the example to a human to label. Then, the classifier is updated with the new labeled example and the process repeats until a suitably accurate classifier has been found.

For the querying step of active learning, there are various measures of how "unsure" a classifier is about a data point. For this project, a multi-class support vector machine (SVM) algorithm was used for classification. The multi-class SVM works by finding a hyperplane in the feature space that separates each pair of classes. These hyperplanes can be used to estimate the probability that each of the events is in one of m different classes [12]. The event selected for querying is the one with the smallest difference between the probability estimates for the two most likely classes. This approach is suggested by prior work [13] because it provides a better indication that the classifier is "confused" about an example than a measure such as entropy, which can be high even when the classifier is relatively confident about the class label.

3.5 Implementation

The techniques described above are implemented in Python using the `NumPy` and `SciPy` libraries [14, 15]. The detection process was performed on the Swinburne supercomputer, and is easily parallelized across individual observation files. A tuple of the observation identifier, start time sample, and end time sample are used to identify an event and describe its extent within the observation. For each event, the data observed during the event is extracted to be processed on local computers.

After features are extracted, there are used as inputs to the unsupervised and active learning algorithms. The Mini-Batch k-means and SVM algorithms are implemented in the `scikit-learn` library [16].

4 Results and Discussion

4.1 General Parkes RFI Trends

Using the techniques described in section 3.2, over 5.3 million RFI events were detected in the Parkes Multibeam Pulsar Survey data. Of the detected events, approximately 3.4 million lasted only a single time sample (125 µs). The average event duration is 15.7 time samples (2 ms).

To investigate the occurrence of RFI at different times of day, the roughly 12,000 files in the dataset are treated as independent and identically distributed (i.i.d.) observations of the RFI environment. For each file, a contamination rate is computed as the percentage of samples that exceed the threshold for RFI detection. The rates for each observation

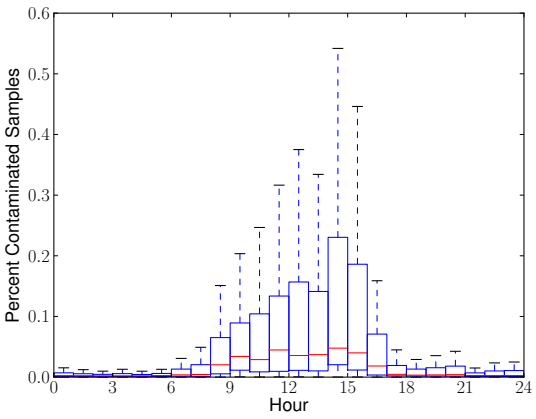


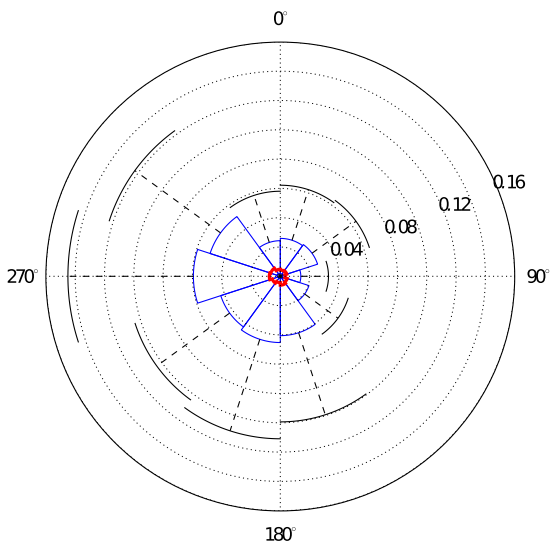
Figure 3: This plot shows the distribution of contamination rates for observations made at different hours of the day. Observations are binned within each hour and treated as independent samples drawn from some underlying distribution. Distributions are represented with box plots showing the median (red), quartiles (blue box), and "whiskers" (dashed line) showing 1.5 times the inter-quartile range.

file are binned into the hour (in local time) in which the observation was made, and the distributions of these rates for each hour are shown as box plots in Figure 3.

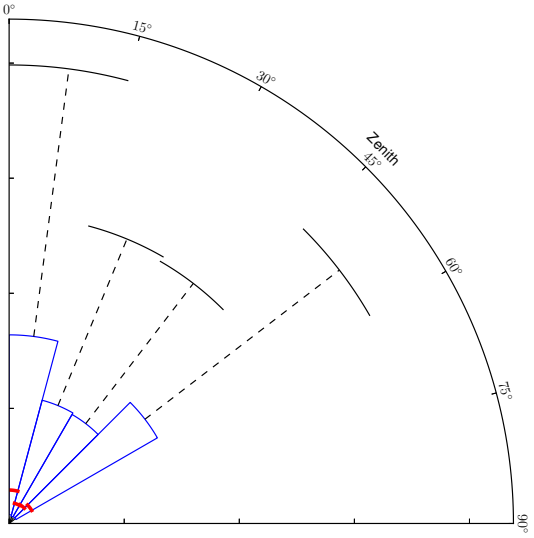
There is an obvious visible difference between the distribution of rates during the work day (approximately 8:00 through 16:00) and the non-work day. Pairwise statistical tests show that the distributions of rates during hours from 8:00 to 16:00 are different than those from 17:00 to 7:00. However, within these blocks of hours, distributions are not significantly different from each other (with only a few exceptions at the edge of the workday). To test the statistical significance of these differences, a Kolmogorov–Smirnov (K–S) test is performed between the distributions corresponding to each pair of hours using an $\alpha = 10^{-18}$ significance level. Note that because these tests compare distributions of rates, they are biased by the fact that rates are only measured during periods of observations, but they are not biased by the fact that observations might be made more frequently at different hours of the day.

A similar analysis can be used to determine whether RFI is observed with different rates as the telescope points in different directions. In Figure 4, binning is performed across azimuth and zenith angles, and the contamination rate is plotted radially outwards. The radial box plots in Figure 4a show the distributions of contamination rates across various azimuth angles, and Figure 4b shows the distribution of rates at different zenith angles. Using a K–S test, the distribution of rates at the highest and lowest zenith angles are similar to each other, but different from the middle zenith angles at an $\alpha = 10^{-3}$ significance level. This makes sense, since at these angles, the receiver is either not blocked by the dish for RFI sources at the observatory, or pointing near the horizon.

To give context for the azimuthal plots, Figure 5 shows a map of the buildings at the Parkes observatory, with Figure 4a in place of the telescope. There are higher rates of RFI observed when the telescope is pointing towards the visitor center, parking lot, café, and generator hut than when it points away from these buildings. These associations are not surprising, but they reinforce the intuition that features such as time of day and telescope pointing direction are useful for characterizing RFI sources.



(a) Distribution of RFI rates across azimuth angles



(b) Distribution of RFI rates across zenith angles

Figure 4: Distributions of contamination rates across various telescope azimuth and zenith angles. Observations are binned within each range of angles and the distributions are plotted as in Figure 3.

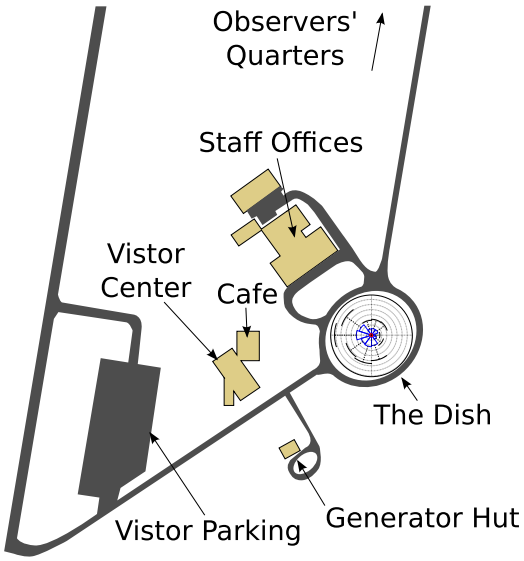


Figure 5: A map of the Parkes observatory, with Figure 4a placed above the telescope dish.

4.2 Clustering Results

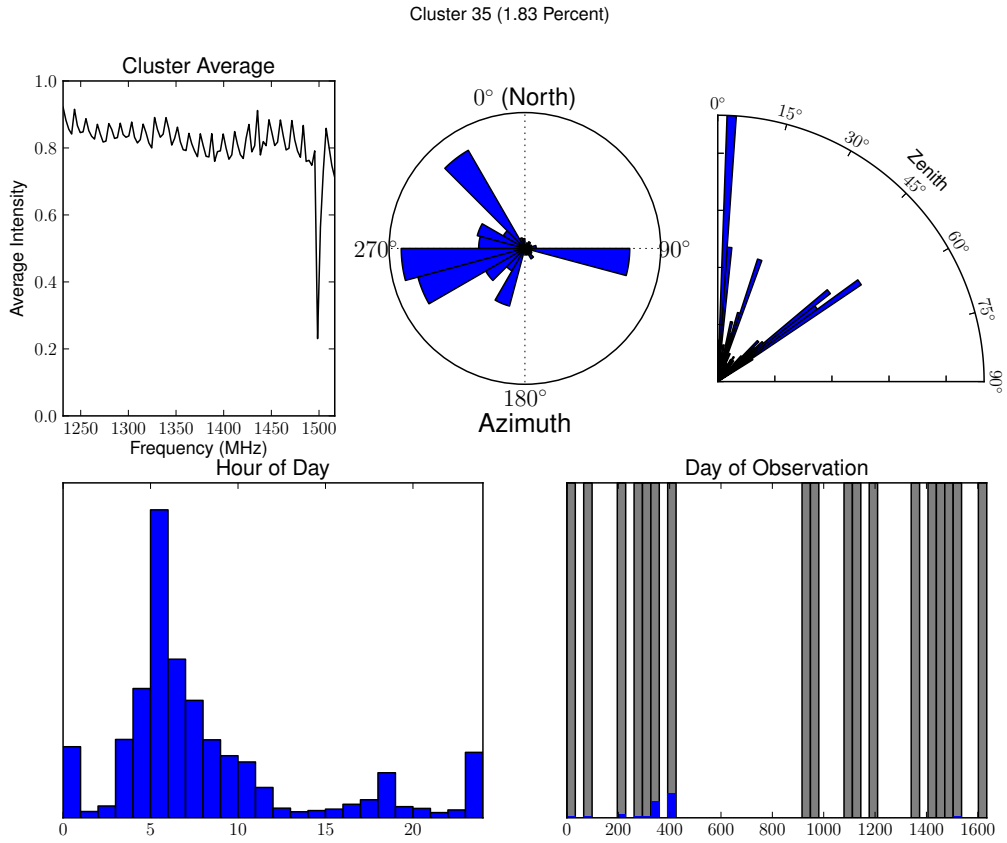


Figure 6: Characteristics of an example cluster, such as the relative frequencies of events when the telescope is pointing in different directions, or observing at various hours of the day.

To eliminate effects of variability in temporal structure across events, the clustering results below use features described in section 3.3 extracted from all events lasting a single time step. As discussed in section 4.1, such events constitute approximately 65% of all detected events.

For Mini-batch k-means, $k = 50$ was chosen to find 50 clusters of related events. A large value of k was chosen to sufficiently separate distinct events sources. With a smaller k value, some distinct sources might be grouped together into the same cluster. However, some expert feedback is necessary to validate or refine the choice of k.

Cluster sizes range from approximately 3% of all events to 1% of events. Cluster sizes can be used to estimate relative and absolute frequencies of events. The characteristics of each cluster are summarized with several graphs. An example cluster is shown in Figure 6. The top left plot shows the cluster center, which is an average across events in the cluster of intensity values at each frequency channel. The top center plot shows the fraction of events at each azimuth angle bin in the cluster relative to all events occurring in that bin. Similarly, the top right plot shows the fraction of events at various zenith angles that are in this cluster. The bottom left plot shows the relative fraction of events within each hour occurring in the cluster. Finally, the bottom right plot shows how the

events occur over the duration of the survey, given in terms of the number of the days since the first observation. The blue portions of the gray bars show the fraction of events in the cluster over days when observations were made. This particular cluster represents nearly 2% of all events, or approximately $60,000$ events.

An expert might be able to interpret cluster statistics to determine a source of the RFI. For example, cluster 35 seems to represent events with high intensity value across all frequency channels except around 1500 MHz. Furthermore, the events seem to occur most often in the morning and evening when the telescope is pointing towards the visitor center parking lot. Therefore, one can speculate that this cluster corresponds to RFI emitted by the spark plugs of arriving and departing vehicles. By visual inspection, approximately 2/3 of clusters exhibit obvious patterns in either average event intensity, time of day, or telescope direction. However, more analysis by an RFI expert is required to say with confidence which clusters correspond to potential RFI sources.

4.3 Active Learning

The clustering results described above group events without any human intervention. However, *some* input by an expert might be able to significantly improve the meaningfulness of grouped events. To collect expert input for RFI classification, an active learning system as described in section 3.4 was developed using a web interface. The web interface allows the user to access the system on a remote computer capable of efficiently processing the millions of detected events. Screen-shots of the interface are shown in Figure 7.

There are currently no significant results to report for the active learning system, as the process of collecting expert labels is ongoing. The labels of 31 events have been collected on a subset of the detections. For these events, the average time to select a label after viewing the selection page (Figure 7a) is about 5 seconds, which corresponds to about 12 labels per minute. Therefore, the active learning system can be used to acquire on the order of hundreds of labels in less than an hour.

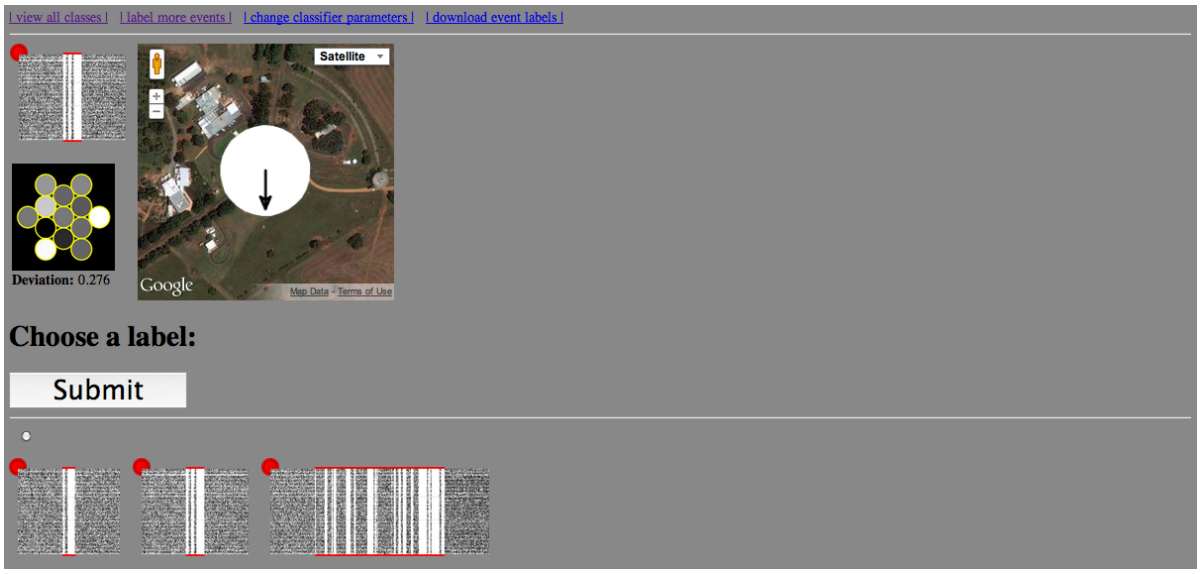
4.4 Future Directions

More expert feedback is necessary to properly interpret cluster results. Similarly, expert use of the active learning system will generate additional classification results. Class labels will not only provide information about possible RFI sources, but inform the development of additional techniques such as feature selection.

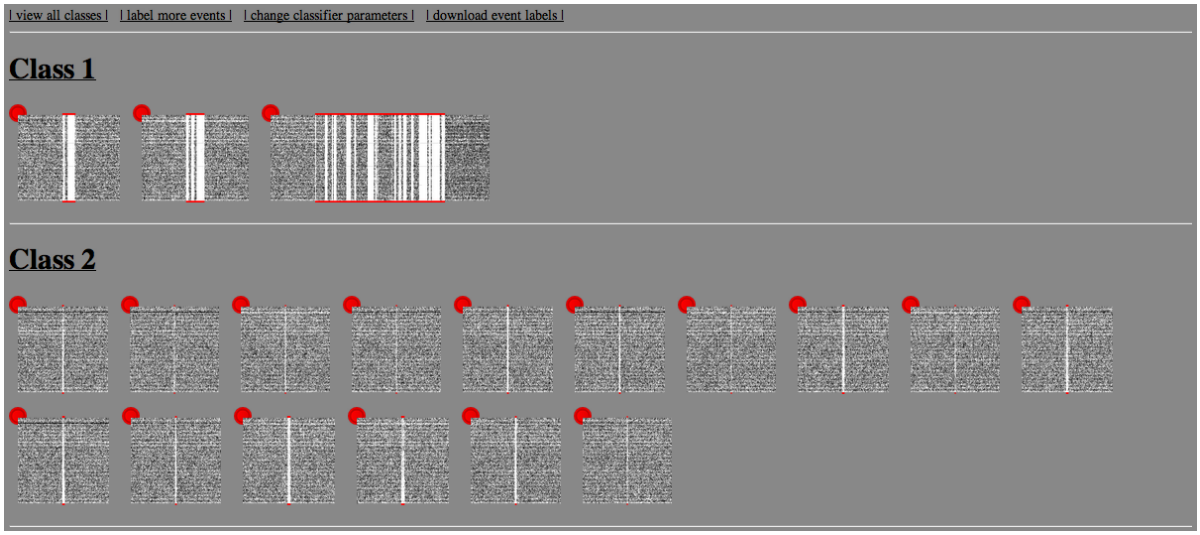
The general framework described above of event detection, feature extraction, and unsupervised or active learning is modular and can be modified in several ways. First, the process can be applied to other instruments and observatories such as the Green Bank telescope or the Deep Space Network (DSN). Furthermore, the event detection algorithm can be used to search for dispersed signals, since some types RFI can exhibit apparent dispersion.

5 Conclusion

Although more expert feedback is needed, this project shows that applying machine learning techniques to detected RFI events has the potential to improve the understanding of various RFI phenomena. Understanding characteristics of RFI provides valuable information that can be used to formulate and prioritize mitigation strategies and optimize



(a) When selecting a label for an event, the user is presented with a set of event statistics and a list of candidate classes.



(b) As events are labeled, the classifier is retrained and the resulting classes are presented to the user.

Figure 7: These screen-shots are taken from the active learning interface for labeling RFI events.

observation scheduling. The techniques described above can generalize to other detection procedures, such as those that look for dispersed signals, to explore other kinds of RFI. A similar characterization can be performed for other instruments and observatories such as the Green Bank or DSN Telescopes. It is my hope that this work informs and motivates an investigation of the application of machine learning approaches to other RFI mitigation problems.

References

[1] Hogden, J.; Wiel, S. V.; Bower, G. C.; Michalak, S.; Siemion, A.; and Werthimer, D.: Comparison of Radio-frequency Interference Mitigation Strategies for Dispersed Pulse Detection. *The Astrophysical Journal*, vol. 747, no. 2, 2012, p. 141.

[2] Fridman, P. A.; and Baan, W. A.: RFI mitigation methods in radio astronomy. *Astronomy & Astrophysics*, vol. 378, no. 1, 2001, pp. 327–344.

[3] Offringa, A. R.; de Bruyn, A. G.; Biehl, M.; Zaroubi, S.; Bernardi, G.; and Pandey, V. N.: Post-correlation radio frequency interference classification methods. *Monthly Notices of the Royal Astronomical Society*, vol. 405, no. 1, 2010, pp. 155–167.

[4] Baan, W. A.: RFI Mitigation in Radio Astronomy. *RFI mitigation workshop*, 2010. URL `http://pos.sissa.it/archive/conferences/107/001/RFI2010_001.pdf`.

[5] Fridman, P. A.: RFI excision using a higher order statistics analysis of the power spectrum. *Astronomy & Astrophysics*, vol. 368, no. 1, 2001, pp. 369–376.

[6] Lyne, A. G.; Camilo, F.; Manchester, R. N.; Bell, J. F.; Kaspi, V. M.; D'Amico, N.; McKay, N. P. F.; Crawford, F.; Morris, D. J.; Sheppard, D. C.; and Stairs, I. H.: The Parkes Multibeam Pulsar Survey: PSR J1811-1736, a pulsar in a highly eccentric binary system. *Monthly Notices of the Royal Astronomical Society*, vol. 312, 2000, pp. 698–702.

[7] Staveley-Smith, L.; Wilson, W. E.; Bird, T. S.; Disney, M. J.; Ekers, R. D.; Freeman, K. C.; Haynes, R. F.; Sinclair, M. W.; Vaile, R. A.; Webster, R. L.; and Wright, A. E.: The Parkes 21 cm multibeam receiver. *Publications Astronomical Society of Australia*, vol. 13, no. 2, 1996, pp. 243–248.

[8] Lyne, A.; and Graham-Smith, F.: *Pulsar Astronomy*. Cambridge Astrophysics Series, Cambridge University Press, 2006.

[9] Bishop, C.: *Pattern Recognition and Machine Learning*. Information Science and Statistics, Springer, 2006.

[10] Sculley, D.: Web-scale k-means clustering. *Proceedings of the 19th International Conference on World Wide Web*, ACM, 2010, pp. 1177–1178.

[11] Cohn, D.; Atlas, L.; and Ladner, R.: Improving Generalization with Active Learning. *Machine Learning*, vol. 15, no. 2, May 1994, pp. 201–221.

[12] Platt, J. C.: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, MIT Press, 1999, pp. 61–74.

[13] Joshi, A.; Porikli, F.; and Papanikolopoulos, N.: Multi-class active learning for image classification. *IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 2372–2379.

[14] Ascher, D.; Dubois, P. F.; Hinsen, K.; Hugunin, J.; and Oliphant, T.: *Numerical Python*. Lawrence Livermore National Laboratory, Livermore, CA, 2001. `http://numpy.scipy.org/`.

[15] Jones, E.; Oliphant, T.; Peterson, P.; et al.: SciPy: Open source scientific tools for Python. 2001–. http://www.scipy.org/.

[16] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and E., D.: Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–2830.